Evaluating the Impacts of Item Exposure Procedures on Ability Estimates in CAT When Items are Disclosed

Wei He

Mark Reckase

Michigan State University

Paper prepared for the 2008 NCME annual meeting. Correspondence concerning this paper should be addressed to Wei He, College of Education, Michigan State University. E-mail: hewei1@msu.edu.

Evaluating the Impacts of Item Exposure Procedures on Ability Estimates in CAT When Items are Disclosed

Objectives of the Inquiry

The purpose of this study is to investigate which item exposure control procedure allows the minimum negative effect on ability estimates in the computerized adaptive test (CAT) given some items in the item pools are disclosed[1]. The disclosed items in this study are defined as items released to the public or intentionally memorized by the test-takers and shared with future examinees.

Theoretical Perspective

Test security has been widely acknowledged as one of the concerns for CAT (eg, Davey, 2007; Yi, et al., 2007; Wainer, 2000; Patsula & Steffen, 1997; Mills & Stocking, 1996) mainly due to its nature of continuous testing. This concern largely motivates the development of exposure control algorithms (see Davey (2007) for a complete review over item exposure procedures). Way (1998) divided item exposure procedures into two categories: randomization and conditional selection procedures. Such procedures as the 5-4-3-2-1 proposed by McBride and Martin (1983) and the randomesque procedure (RS) (Kingsbury & Zara, 1985) belong to the first category. The most fundamental, perhaps also the most commonly-used conditional selection procedure, is the Sympson-Hetter (SH) procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985). Based on the SH, a series of other conditional selection procedures were developed, such as the Davey and Parshall (DP) procedure (Davey & Parshall, 1995; Parshall et al., 1998), the Stocking and Lewis unconditional multinomial (SL) procedure (Stocking & Lewis, 1995), and the Stocking and Lewis conditional multinomial (SLC) procedure (Stocking & Lewis, 1998). In addition, Chang and Ying (1999) proposed *a*-stratified procedure and indicated that this procedure can satisfactorily control the item exposure by better balancing item use rates. Later, this procedure has been further explored by incorporating more elements such as the SH procedure and content balancing (Chang, et al., 2001; Leung, et al., 2003).

[1] Item disclosure/theft/memorization are used interexchangeably in this proposal.

Several studies (Boyd, et al., 2002, Chang, 1998; Chang et al., 2000; Davey & Parshall, 1995; Chang, et al., 2003, Revuelta & Ponsoda, 1998) have been conducted to evaluate the performance of different item exposure strategies. Chang and Ansley (2003) systematically compared the performance of five exposure control algorithms. They reported that the SLC procedure best serves the purposes of controlling the observed exposure rates to the desired values as well as producing the lowest test overlap rate. In addition, they also reported the trade-offs between item exposure and measurement precision and similar findings can be found in other studies (Boyd, et al., 2003).

If the Kaplan event[2] triggered the war on security maintenance of continuous CAT between testing organizations and examinees/coaching schools, this war to date appears so unprecedentedly tough to win by the testing organizations, especially with the increasingly easy access to the World-Wide-Web (WWW) throughout the world. With WWW, some examinees routinely post on the internet their recollections of items they were administered, leaving the future examinees with clear opportunities to take advantage of previous exposure to some or many of the questions, subsequently inflating their test scores. Measurement professionals respond by taking a variety of approaches, for example: developing indices to detect the items that are compromised due to the pre-knowledge of the items, eg. l_z, ECI4z, and ECI6z indices (See Meijer, 1996, for a review), researching methods on how to make good use of recycled items (Chang, 2007; Yi, et.al, 2007), or investigating the impact of memorized items on the test performance (Schipke & Scrams, 1999; Stocking, et al., 1998). Schnipke and Scrams (1999), by simulating item theft by regular and professional test thieves, reported that examinees, especially the low-ability examinees, can receive ability estimates at the top of the ability range when stolen items were provided by professional thieves. Similar finding was reported by Yi, et al. (2007). However, Stocking et al. (1998) indicated that use of questions previously made available to the public has effects on test scores small enough to be acceptable.

[2] This refers to what Kaplan Educational Centers did to CAT GRE in 1994 by registering its employees to take the test and memorize the questions. The memorized items were then collected by Kaplan to reconstruct a large portion of item pool on which the test was based.

By and large, the widespread concerns over CAT security does not prevent CAT from being an appealing measurement tool due to its unrivaled advantages over traditional paper-and-pencil tests in terms of time efficiency, management flexibility, measurement precision, and timely score reporting, and so on. It is used in a large number of large-scale operational tests, for example, AICPA and NCLEX exams. That is to say, although unwillingly and unhappily, measurement professionals still have to live with the fact—hopefully just for a while—that some examinees may have acquired the pre-knowledge of test items that appear in their tests, and, as a result, their ability estimates may be inflated. Considering different CAT programs may use different item exposure procedures, it is necessary to conduct a systematic research investigating the impacts of different item exposure procedures on the ability estimates at the presence of the disclosed items.

Therefore, the purpose of this study is to examine which item exposure control procedure allows the minimum negative effect on ability estimates in the computerized adaptive test given some items in the item pools are disclosed. Specifically, the following two research questions were addressed:

1) To what degree are ability estimates inflated at the presence of disclosed items?
2) Given ability estimates are inflated, which item exposure control procedure can allow the smallest negative effect on ability estimates?

Method

Monte Carlo simulation was used in this study. Matlab (V7.0) was used for the simulation.

Item Exposure Control Procedures

Four procedures were considered in this study: 1) the randomesque procedure (Kingsbury & Zara, 1985) which involves selecting one of five items that can maximize the item information at the current ability estimate; 2) the SH procedure (Hetter & Sympson, 1997); 3) the *a*-stratified procedure (AS) (Chang & Ying, 1999); and 4) the a-stratified with *b*-blocking (BAS) procedure (Chang, Qian, & Ying, 2001). To facilitate the

interpretation, a CAT procedure implementing no item exposure control (WO) was also considered in this study.

Item Pool Characteristics

The item pool used in this study comes from a large-scale operational CAT program. This item pool contains 270 items. Table 1 gives overall item parameter descriptive statistics of this item pool. The correlation between *a*- and *b*-parameters is .556.

[Insert Table 1 about here]

To implement the AS and BAS procedures, the item pool had to be stratified. The item pool was divided into 5 strata with equal size of 54 items in each stratum. A detailed description on how to partition the item pool in AS and BAS structures can be referred to Chang and Ying (1999) and Chang et al. (2001) respectively. Table 2 presents the descriptive statistics of AS and BAS item pools across different stratum. It can be observed that both a- and b-parameters in the AS item pool increase in magnitude across different stratum, while for the BAS item pool only a-parameter increases across different stratum with b-parameter remaining constant.

[Insert Table 2 about here]

Simulating Disclosed Items

The features of disclosed items are expected to affect final ability estimates in a different manner due to the adaptive nature of item selection implemented by the CAT. For example, it is very likely that low-ability examinees may have slim chance to see very difficult items, or vice versa. Therefore, how to simulate disclosed items should be well considered so that these disclosed items can be close as much as possible to what they are supposed to be in the real life and research findings can be informative. Prior studies of relevant topic took two approaches in this regard. One was to randomly sample a certain percentage of items and then treat them as disclosed items (e.g., Stocking et al., 1998). This approach was justified by the way that a testing company may release the items. For example, the testing company may need to periodically publicize certain number of items to ensure its integrity and credibility. In this case, it may be reasonable to assume that the

features of disclosed items may be representative of items contained in the operational item pool. The other was to run a certain CAT algorithm and treat those items with the highest exposure rates as disclosed items (Ying, et al., 2007; Schnipke and Scrams, 1999). This approach was justified by the fact that items seen most by the examinees were very likely to be items on which future examinees might have better knowledge given that they are shared with the future examinees. This study took both approaches into consideration and identified two sets of disclosed items. The first set (called RIS_set) was a random sample of 10% (i.e., 27) items from the item pool. To identify the second set, five CAT procedures as described below were run using the same 25,000 examinees randomly selected from the standard normal distribution. Then, the counts of item exposure of each item from each CAT procedure were tallied. Based on the magnitude of the counts, items were rank-ordered from the most to the least exposed and those at the top 10% (i.e., 27 items) were considered as disclosed. Table 3 presents the descriptive statistics of these five sets of disclosed items. It is within our expectation that the CAT program implementing no item exposure control procedure yielded a set of items with the highest average item discrimination parameter, followed by RS and SH procedures. Interestingly, the item set yielded by the stratified item pool had the highest average item difficulty but lowest average item discrimination parameters. The item set yielded by the SH procedure had the medium-level average item discrimination and difficulty parameters. Considering the widespread concern that the maximum item selection method, which is often used in the operational CAT program, tends to select an item with high item discrimination parameter, the item set yielded by no exposure control procedure was picked as the second set (called MI_set) disclosed items in our study. Table 4 compares the characteristics of these two sets of disclosed items. Obviously, the items in the MI_set are more difficult than those in the RIS_set. Tables 5 and 6 present both number and characteristics of these two sets of disclosed items across different stratum of the item pool structured according to AS and BAS designs. When the RS_set was used, the disclosed items were somewhat evenly distributed in each stratum in both AS and BAS designs. However, these disclosed items were found only in the last three strata when the MI_set was used. As indicated by Table 6, a large number of disclosed items were contained in the last two strata.

[Insert Table 3 about here]

[Insert Table 4 about here]

[Insert Table 5 about here]

[Insert Table 6 about here]

CAT Components

The three-parameter logistic item response model (3-PL) (Birnbaum, 1968) was used as the IRT model in this study. For the WO, the RS, and the SH procedures, item selection method is maximum item information selection. For both AS and BAS designs, the item that can minimize the absolute difference between b and $\hat{\theta}$ was selected from the stratum corresponding to the test stage. A combination of Owen's Bayesian procedure (1975) and maximum likelihood estimation (MLE) was used for ability estimation. For the Owen's procedure, a normal prior (0,1) was used. Once both correct and incorrect responses were available, MLE was used. All examinees were assumed with the abilities of zero upon the starting of the test. Test length was fixed at 25. For both AS and BAS design, five items were selected out of each strata respectively. No content balancing was considered in this study.

Procedures for Data Simulation

For each CAT employing different item exposure control procedure, a series of simulation was carried out under the following stages:

Stage 1: Determine item exposure parameters for the SH procedure. The target maximum-desired exposure rate was set at .2, which was considered as a reasonable rate for most operational CAT programs (e.g, Eignor, Stocking, Way, & Steffen, 1993; Schaeffer et al., 1995; Chang & Ansley, 2003). A sample of 25,000 examinees drawn from the standard normal distribution was used to generate the item exposure parameter. A total number of 16 runs were conducted. After the 12^{th} iteration, the observed maximums were stabilized.

Stage 2: Run the CAT algorithm for each simulee with item pool containing disclosed items. To better understand the impacts of item exposure procedures on the ability estimates at the presence of disclosed items, a conditional sample of

3,500 examinees at each of the ability levels equally spaced between -3 to 3 at an interval of .5 (i.e., -3, -2.5,…2.5, 3, totaling 13 discrete ability points) were administered the CAT implementing different item control procedures. When a simulee was administered a disclosed item, a correct response was automatically assigned regardless of ability level. The average number of disclosed items administered conditional on each discrete theta point was tallied.

Stage 3: Run the CAT algorithm to the same simulees as those in Stage 2 with the item pool containing none of the items assumed as disclosed. These ability estimates were to be used for the paired-sample t-tests with the purpose of investigating the impacts of disclosed items on final ability estimates.

Analysis

The following steps were undertaken to analyze the simulation results. To start with, the final ability estimates yielded by the item pool containing disclosed items and no disclosed items for each discrete ability point were compared with each other across different item exposure control procedures in light of 1) conditional bias (Eq. 1); and 2) conditional mean square error (MSE) (Eq. 2). Secondly, paired-sample t-tests were conducted on each discrete conditional ability point to investigate whether the presence of disclosed items statistically affected the final ability estimates given by different item exposure control procedures. Effect size was also reported. Finally, the multivariate analysis of covariance (MANCOVA) was conducted to investigate the impacts of item exposure control procedures on the final ability estimate accuracy and precision, controlling the difference between true ability and estimated ability estimates when there were no disclosed items as covariate and treating bias and MSE as dependent variables.

$$\text{Conditional bias} = \frac{1}{N_i}\sum_{j=1}^{N_i}(\hat{\theta}_{j_i} - \theta_i) \qquad \text{Eq.1}$$

$$\text{Conditional MSE} = \frac{1}{N_i}\sum_{j=1}^{N_i}(\hat{\theta}_{j_i} - \theta_i)^2 \qquad \text{Eq.2}$$

where θ_i = -3, -2.5, …, 2.5, 3 for i = 1, 2, …, 13, respectively, and $\hat{\theta}_{j_i}$ (j= 1, 2, …, 3500) is the estimator of θ .

Results

Average number of disclosed items administered

Tables 7 and 8 report the average number of disclosed items administered conditional on each theta point for both RIS_set and MI_set disclosed items. It can be observed that, for the RIS_set disclosed items, the average number of disclosed items administered to examinees with abilities conditional on different theta points somewhat followed the same pattern across different item exposure control procedures with low-ability examinees being administered more disclosed items than the others. However, for the MI_set disclosed items, the RS, AS, and WO yielded considerably large number of disclosed items especially for those examinees with true abilities below -.5. BAS yielded the smallest number of disclosed items.

[Insert Table 7 about here]

[Insert Table 8 about here]

Conditional Bias

Panels A and C in Figure 1 depict the conditional biases given by the item pools containing RIS_set and MI_set disclosed items respectively. Panel B, which portrays the conditional biases yielded by the item pool containing no disclosed items, indicates that different exposure procedures yielded comparable biases for ability points from -2 to 2. As the result of presence of disclosed items in the item pool, however, the magnitudes of biases for the ability points lower than .5 started to exhibit noticeable increase, suggesting that disclosed items tended to inflate the ability estimates of these examinees. Specifically, the change of conditional biases followed the same trend across different item exposure procedures for the RIS_set disclosed items. First of all, the lower the true ability was, the more increase the bias tended to have. Second, a systematic increase of bias could be observed across different item exposure control procedures. For the MI_set disclosed items, it can be noted that the CAT algorithm using the RS procedure behaved in the same way as that implementing no item exposure control procedure. For those

examinees whose true abilities were lower than -1, their final ability estimates can be inflated by at least 1.5 by the RS. Nonetheless, this magnitude was just approximately from .4 to .8 by the other three exposure control procedures including SH, AS, and BAS. By and large, the MI_set disclosed items produced different patterns of the change of biases across all exposure control procedures with the BAS procedure appearing to inflate the final ability estimates by the least amount.

[Insert Figure 1 about here]

Conditional MSE

Panels A and C in Figure 2 depict the conditional MSEs given by the item pools containing RIS_set and MI_set disclosed items respectively. Panel B, which portrays the conditional MSEs given by the item pool containing no disclosed items, indicates that, in general, measurement precision for those examinees with true abilities above 0 was better than those with true abilities below 0. In addition, the RS procedure appeared to produce the lowest MSEs along the whole proficiency scale while it was the SH procedure that yielded the largest conditional MSEs along the whole proficiency scale. This pattern remained unchanged even at the presence of RIS_set disclosed items except the increase of the magnitude of MSEs, which suggested the presence of RIS_set did negatively impact the measurement precision, especially for those examinees whose true abilities are below -2. The MI_set disclosed items, as demonstrated by Panel C, resulted in larger increase of conditional MSEs than the RIS_set. For this set, the BAS procedure appeared to cause the least change of conditional MSEs among all item exposure control procedures under study and the RS procedure increased the MSEs by the most substantial magnitude.

[Insert Figure 2 about here]

Paired-sample t-tests

Both Tables 9 and 10 report the paired-sample t-test statistics for different discrete ability points and their corresponding effect sizes. Effect sizes are categorized as “small”,

"medium", "large", and "extremely large" and highlighted in different manner. As can be observed from these two tables, both RIS_set and MI_set disclosed items tended to benefit low-ability or even medium-ability examinees and affect no high-ability examinees. The MI_set disclosed items were found to inflate final ability estimates much more than the RIS_set disclosed items by having much higher effect sizes almost across all item exposure control procedures. However, it seems that the RIS_set might affect more examinees than the MI_set disclosed items. The MI_set appeared to exert impacts on those examinees with true abilities below 1, while for the RIS_set, those examinees with true abilities below 1.5 appeared affected. Figure 3 graphically depicts the effect sizes associated with different sets of disclosed items. Clearly, these two sets of disclosed items produce different patterns of effect sizes along the whole proficiency scale. For the effect sizes caused by the RIS_set, there appeared to be certain degree of interaction between different conditional theta points and exposure control procedures. For the effect sizes caused by the MI_set, both RS and WO provided the worst results.

[Insert Table 9 about here]

[Insert Table 10 about here]

[Insert Figure 3 about here]

MANCOVA Results[3]

Tables 11 to 14 report the MANCOVA results when the RIS_set was used as disclosed items. As indicated by Table 11, the interaction between ExpoCtrl (i.e., main factor—item exposure control procedures) and DIFF (i.e., covariate—deviation of conditional ability point estimate from true ability) was not statistically significant. Therefore, this term was dropped out of the model. Table 12, which describes the multivariate tests for the main effect, suggests that exposure control procedure was statistically significant with $F=2.348$[4] and $p=.021$. This significance suggests different item control procedure could affect final ability estimate accuracy and precision in different ways when difference between true ability and ability estimates given by the item pool containing no disclosed

[3] For the MANCOVA, a sample size of 75 (15 ability points*5 levels) was used for the consideration of statistical power. However, only 13 ability points were used in the previous analyses.

[4] We referred to Wilks' Lambda because five groups were involved in our analysis.

items was controlled. Table 13 further shows that different item exposure control procedures may cause statistically significant difference in bias with df=4, F=3.796, and p=.008. The pairwise comparison with Bonferroni adjustment, as suggested by Table 14, suggests that the BAS yielded the smallest bias—0.042 less than that from either of RS, SH, and AS procedures. What's more, the biases yielded by the RS, SH, and AS procedures were not statistically different from each other.

[Insert Table 11 about here]

[Insert Table 12 about here]

[Insert Table 13 about here]

[Insert Table 14 about here]

Tables 15 to 18 report the MANCOVA results when the MI_set was used. Unlike the results from the RIS_set, the interaction between ExpoCtrl (i.e., main effect) and DIFF (i.e., covariate) was statistically significant with F=14.137 and p<.001. This significance made main effect hard to interpret because main effect confounded with covariate. Table 16, which describes the between-subject effect, indicates that bias was significantly different across different exposure control procedures with df=4, F=3.242, and p=.017. Pairwise comparisons for bias and MSE, as demonstrated by Table 17 and Table 18 respectively, indicate that among all item exposure control procedures the BAS may produce the least bias and MSE. That is to say, given that item disclosure can significantly inflate ability estimates, the BAS procedure can allow the minimum negative effects on ability estimate accuracy and precision. However, cautions should be exercised to interpret the main effect due to the statistical significance of interaction term,.

[Insert Table 15 about here]

[Insert Table 16 about here]

[Insert Table 17 about here]

[Insert Table 18 about here]

Conclusion and Discussion

Simulating the use of disclosed items in the contexts of CATs implementing four different item exposure control procedures and no item exposure control procedure, this study investigated which procedure might allow the minimum negative impacts on final ability estimate accuracy. Two sets of disclosed items with total number of 27 (i.e., 10% disclosure rate) in each set were simulated under different rationale. The results suggested that features of disclosed items affected final ability estimates in different ways. Specifically, the RS_set disclosed items yielded different patterns of change of bias and MSE from the MI_set. Using the RS_set, both biases and MSEs witnessed a systematic increase across different exposure control procedures especially for the low-ability examinees with the true abilities lower than -1.5. Using the MI_set, however, both biases and MSE displayed different patterns of change across different exposure control procedures with the RS and the BAS procedures causing the largest and the smallest increase respectively. Despite the differences of two sets of disclosed items, they were found to be able to significantly inflate the final ability estimates especially for those low-ability examinees, i.e. those with true abilities below -2. What's more, the MI_set disclosed items were found to inflate final ability estimates much more severely than the RS_set. The MANCOVA revealed that, after controlling the difference between the true ability and the conditional ability estimates given by the item pool containing no disclosed item, the BAS procedure performed the best among all these procedures under study at the presence of disclosed items by having the smallest bias. The rest of other procedures did not perform statistically different from each other with regard to bias. The magnitude of MSEs yielded by different procedures did not statistically differ from each other. However, cautions should be taken to generalize this finding since for the MI_set a statistically significant interaction was detected. As expected, the CAT implementing no item exposure control procedure performed the worst.

The result suggests that different item exposure procedures work differently at the presence of disclosed items. Among all item exposure procedures under study, the BAS performed the best. Compared with other procedures including the RS and the SH

procedures which involve no stratification of item pools, the stratification design requires to fix number of items to be selected out of each stratum, therefore reducing the rate that a certain disclosed item can be selected. However, although the AS pool was also structured for the same consideration, failure to take into the correlation between the a- and b-parameters—unlike what the BAS does—may explain the difference in ability estimates accuracy and precision from the BAS when an item pool contain disclosed items.

The results of this study are in consistent with Schnipke & Scrams (1999) in terms that the presence of disclosed items can inflate the final ability estimates in the context of CATs by an unacceptably high degree, especially for the low-ability examinees. However, the similar study by Stocking et al. (1998) indicated that using up to about 10% disclosed items in CATs produced increases in observed tests scores that may be viewed as small when compared with changes in test scores from retesting. More research may be needed in this issue, especially on how different features of disclosed items may impact final ability estimates differently. Future studies in this regard should also include more exposure control procedures and investigate how different sizes of disclosed items impact final ability estimates across different item exposure control procedures.

References

Chang, H-H, Zhang, J., & Yi, Q. (2007). Some empirical results concerning how to use "recycled" items in computerized adaptive testing. presented at the Annul Meeting of the National Council of Measurement on Education. Chicago, IL.

Chang, H-H, & Ying, Z. (1999). A-stratified multistage in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222.

Chang, H., Qing, J., & Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25, 333-341.

Chang, S-W, & Ansley, T.N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. Journal of educational measurement, 40(1), 71-103

Chang, S-W, Ansley, T.N, & Lin, S-H (2000). Performance of Item Exposure Control Methods in Computerized Adaptive Testing: Further Exploration. Paper presented at the Annul Meeting of the American Educational Research Association. New Orleans, LA April.

Davey, T. & Stone, E., (2007). Improving security under continuous testing. Paper presented at the Annul Meeting of the National Council of Measurement on Education. Chicago, IL.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (eds.), *Computer-Based Testing: Building the Foundation for Future Assessments* . Mahwah, NJ: Lawrence Erlbaum.

Hetter, R., & Sympson, B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. Wasters, & J. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.

Leung Chi-keung, Chang, H-H, Hau, K-T. (2003). Incorporation of Content Balancing Requirements in Stratification Design for Computerized Adaptive Testing. *Educational and Psychological Measurement, 63*(2), 257-270.

McLeod, L.D., & Lewis C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, 23(2), 147-160.

Meijer, R.R. (Ed.). (1996). Person-fit research: Theory and applications [Special issue]. *Applied Measurement in Education*, 9(1), 9-18.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of educational Measurement*, 35(4), 311-327.

Schnipke, D. L & Scrams, D.J, (1999). Item theft in a continuous-testing environment: what is the extent of the danger? (Computerized Testing Report 98-01). Newtown, PA: Law School Admission Council.

Stocking, M. L., Ward, W.C., & Potenza, M.T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of educational measurement,* 35(1), 48-69.

Stocking, M.L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics,* 23(1), 57-75.

Sympson, J.B., & Hetter, R.D. (1985, October). *Controlling Item-Exposure Rates in Computerized Adaptive Testing.* Paper presented at the Military Testing Association, San Diego, CA.

Yi, Q, Zhang, J, & Chang, H. (2007). The effects of item pool size on the severity of possible test security violations in CAT. Paper presented at the Annul Meeting of the American Educational Research Association. Chicago, IL.

Appendix

Table 1

Descriptive Statistics of Item Parameters

Item parameter	Mean	Std	Max	Min
a	1.3568	0.4421	2.7534	0.5644
b	0.1369	1.1343	2.3430	-2.6252
c	0.2030	0.0664	0.4592	0.0380

Table 2

AS and BAS Item Pool Statistics

		Level 1		Level 2		Level 3		Level 4		Level 5	
Item parameter	Item pool	AS	BAS	AS	BAS	AS	BAS	AS	BAS	AS	BAS
No. Items	270	54	54	54	54	54	54	54	54	54	54
a											
Mean		0.837	1.016	1.079	1.176	1.286	1.319	1.533	1.514	2.049	1.759
SD		0.105	0.228	0.057	0.305	0.058	0.337	0.098	0.407	0.294	0.475
Max		0.977	1.409	1.173	1.935	1.386	2.168	1.738	2.476	2.753	2.753
Min		0.564	0.564	0.983	0.698	1.175	0.726	1.386	0.839	1.744	1.018
b											
Mean		-1.087	0.131	0.017	0.139	0.344	0.138	0.379	0.137	1.032	0.140
SD		0.923	1.136	1.257	1.129	0.918	1.154	0.751	1.148	0.483	1.147
Max		1.838	2.173	2.343	2.215	2.242	2.343	2.118	2.242	2.211	2.211
Min		-2.538	-2.211	-2.625	-2.201	-1.761	-2.538	-1.334	-2.625	-0.069	-2.355
c											
Mean		0.174	0.200	0.203	0.198	0.224	0.190	0.205	0.209	0.210	0.217
SD		0.052	0.059	0.062	0.064	0.066	0.064	0.063	0.078	0.078	0.065
Max		0.293	0.349	0.396	0.396	0.435	0.323	0.298	0.459	0.459	0.348
Min		0.071	0.071	0.058	0.065	0.093	0.058	0.038	0.038	-0.069	0.052

Table 3

Descriptive statistics of different sets of disclosed items yielded by different item exposure control procedures

Item parameter	No. Item	Exposure control procedures				
		RS	SH	AS	BAS	WO
a	27					
Mean		1.692	1.471	1.236	1.142	1.714
SD		0.361	0.269	0.357	0.305	0.380
Max		2.451	2.100	1.405	1.745	2.746
Min		1.183	0.898	0.852	0.564	1.240
b						
Mean		-0.114	-0.057	0.058	-0.022	-0.124
SD		0.615	0.248	0.823	0.739	0.646
Max		0.977	0.547	1.838	1.335	1.081
Min		-1.108	-0.390	-1.108	-1.566	-1.334
c						
Mean		0.170	0.165	0.194	0.199	0.177
SD		0.056	0.051	0.068	0.056	0.058
Max		0.283	0.270	0.256	0.288	0.283
Min		0.083	0.083	0.138	0.083	0.083

Note. WO=without exposure control procedure

Table 4

Item characteristics of two sets of disclosed items

	No. Items	Item parameter	Mean	SD	Max	Min
MI_set	*27*	*a*	1.714	0.380	2.746	1.240
RIS_set	*27*	*a*	1.245	0.439	2.397	0.654
MI_set	*27*	*b*	-0.124	0.646	1.081	-1.334
RIS_set	*27*	*b*	-0.255	1.256	1.793	-2.355
MI_set	*27*	*c*	0.177	0.058	0.283	0.083
RIS_set	*27*	*c*	0.213	0.075	0.459	0.103

Table 5

Item characteristics for the RIS_set

Item parameter	Item pool	Level 1		Level 2		Level 3		Level 4		Level 5	
		AS	BAS	AS	BAS	AS	BAS	AS	BAS	AS	BAS
No. Items	27	9	6	5	5	4	4	4	6	5	6
a											
Mean		0.811	0.958	1.104	0.995	1.251	1.250	1.499	1.469	1.959	1.513
SD		0.114	0.300	0.051	0.219	0.042	0.244	0.100	0.494	0.252	0.534
Max		0.958	1.290	1.169	1.228	1.290	1.459	1.647	1.925	2.397	2.397
Min		0.654	0.654	1.048	0.723	1.204	0.941	1.431	0.839	1.778	1.048
b											
Mean		-1.243	0.000	-0.924	-0.571	0.392	0.016	0.430	0.155	1.125	-0.839
SD		0.512	1.051	1.593	0.989	0.746	1.054	0.624	1.710	0.677	1.371
Max		-0.627	1.514	1.514	1.003	1.063	1.147	1.147	1.793	1.793	1.190
Min		-2.119	-1.033	-2.355	-1.566	-0.368	-1.309	-0.375	-2.119	0.027	-2.355
c											
Mean		0.183	0.192	0.175	0.236	0.240	0.218	0.208	0.219	0.287	0.204
SD		0.049	0.045	0.043	0.054	0.033	0.062	0.081	0.124	0.113	0.078
Max		0.293	0.246	0.226	0.293	0.272	0.281	0.281	0.459	0.459	0.348
Min		0.136	0.136	0.124	0.154	0.194	0.164	0.103	0.103	0.202	0.124

Table 6

Item characteristics for the MI_set

Item parameter	Item pool	Level 1		Level 2		Level 3		Level 4		Level 5	
		AS	BAS	AS	BAS	AS	BAS	AS	BAS	AS	BAS
No. Items	27	0	0	0	1	4	3	13	8	10	15
a											
Mean						1.276	1.585	1.542	2.020	2.114	2.479
SD						0.035	0.308	0.089	0.264	0.318	0.254
Max						1.306	1.859	1.659	2.272	2.746	2.746
Min						1.240	1.252	1.405	1.745	1.745	2.241
b											
Mean						-0.743	0.104	-0.415	0.549	0.502	0.796
SD						0.269	0.562	0.498	0.298	0.348	0.305
Max						-0.529	0.547	0.295	0.849	1.081	1.081
Min						-1.108	-0.529	-1.334	0.252	-0.068	0.475
c											
Mean						0.146	0.123	0.204	0.169	0.155	0.179
SD						0.039	0.034	0.061	0.066	0.048	0.022
Max						0.197	0.161	0.283	0.225	0.225	0.204
Min						0.112	0.096	0.083	0.097	0.096	0.163

Note. Blank shows no disclosed items in these strata.

Table 7

Average number of disclosed items administered (RIS_set)

T	RS		SH		AS		BAS		WO	
	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
-3	4	0~7	3	0~9	3	0~6	6	0~9	4	1~6
-2.5	4	0~7	3	0~9	3	0~6	5	0~9	4	1~6
-2	4	0~7	3	0~8	3	0~6	5	0~9	4	1~6
-1.5	3	0~7	3	0~8	3	0~6	3	0~9	3	1~6
-1	2	0~6	3	0~7	2	0~6	2	0~7	2	1~6
-0.5	1	0~4	2	0~7	2	0~6	2	0~6	1	1~5
0	2	0~3	2	0~6	2	0~6	2	0~6	2	0~3
0.5	1	0~3	2	0~6	1	0~5	2	0~5	1	0~3
1	2	0~3	2	0~6	1	0~6	2	0~5	2	0~3
1.5	2	0~4	2	0~7	0	0~2	2	0~5	2	1~3
2	2	0~3	2	0~8	0	0~1	0	0~5	2	0~3
2.5	1	0~3	2	0~7	0	0	0	0~3	1	0~3
3	1	0~3	2	0~7	0	0	0	0~1	0	0~2

Table 8

Average number of disclosed items administered (MI_set)

T	RS		SH		AS		BAS		WO	
	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
-3	9	3~13	5	0~10	9	0~11	0	0~8	9	5~11
-2.5	9	3~13	5	0~10	9	0~11	0	0~7	9	4~11
-2	9	3~13	5	0~10	9	0~11	1	0~8	9	5~11
-1.5	9	3~13	5	0~10	9	2~11	2	0~7	9	5~11
-1	9	3~13	5	0~10	8	2~12	4	0~8	9	4~11
-0.5	9	3~13	5	0~9	7	0~12	4	0~8	9	5~11
0	8	4~12	4	0~9	5	0~11	4	0~8	8	5~11
0.5	8	3~13	3	0~10	3	0~9	3	0~7	7	4~11
1	5	2~8	3	0~7	1	0~9	2	0~6	5	4~8
1.5	3	2~6	2	0~6	0	0~4	0	0~6	4	4~5
2	3	2~6	2	0~6	0	0~1	0	0~2	4	4~4
2.5	3	2~4	2	0~5	0	0	0	0~3	4	4~4
3	3	2~4	2	0~5	2	0~5	0	0~1	4	4~4

Table 9

Paired-sample t-tests statistics and effect sizes (RIS_set)

T	**RS**		**SH**		**AS**		**BAS**		**WO**	
	TS	EF	TS	EF	TS	EF	TS	EF	TS	EF
3	-1.22	0.029	0.43	0.007	-5.20*	0.088	0.01	0	0.46	0.008
2.5	2.03**	0.003	3.73***	0.063	-0.28	0.005	0.76	0.013	1.09	0.018
2	9.31***	**0.104**	3.78***	0.064	1.53	0.026	1.58	0.027	8.45***	**0.143**
1.5	10.01***	**0.102**	8.67***	**0.146**	0.5	0.009	8.628***	**0.146**	8.72***	**0.147**
1	26.27***	**0.224**	10.37***	**0.175**	11.00***	**0.186**	13.47***	**0.228**	12.86***	**0.217**
0.5	32.41***	**0.139**	12.95***	**0.219**	10.56***	**0.179**	9.68***	**0.164**	6.75***	**0.114**
0	40.35***	**0.164**	11.56***	**0.195**	12.04***	**0.204**	12.54***	**0.212**	13.82***	**0.234**
-0.5	44.19***	**0.153**	15.92***	**0.269**	18.04***	***0.305***	19.26***	***0.326***	12.58***	**0.213**
-1	41.83***	**0.167**	20.22***	***0.342***	13.28***	***0.225***	17.17***	**0.29**	13.43***	**0.227**
-1.5	47.04***	***0.335***	22.40***	***0.379***	14.77***	***0.25***	27.63***	***0.467***	23.43***	***0.396***
-2	67.87***	***0.59***	25.83***	***0.437***	25.49***	***0.431***	46.50***	**0.786**	35.16***	**0.594**
-2.5	89.42***	**0.735**	32.89***	***0.556***	37.73***	**0.638**	59.12***	**0.999**	47.05***	**0.795**
-3	104.57***	**0.839**	33.91***	***0.573***	51.45***	**0.87**	66.79***	***1.129***	49.91***	**0.844**

Note. T=true ability TS=test statistics EF=effect size
p*<.05 p**<.01 p***<.001
. Bold number indicates small effect size (.1~.29);
Bold+Italicized (in red) number indicates medium effect size (.3~.59);
Bold+underlined number indicates large effect size (.6~.89);
Bold+Italicized+underlined number indicates extremely large effect size (>.9).

Table 10

Paired-sample t-tests statistics and effect sizes (MI_set)

T	RS		SH		AS		BAS		WO	
	TS	EF	TS	EF	TS	EF	TS	EF	TS	EF
3	0.61	0.01	-0.48	0.008	-73.56***	1.243	0.19	0.003	-0.7	0.012
2.5	-1.03	0.017	1.27	0.021	0.35	0.006	0.72	0.012	0.73	0.012
2	0.41	0.007	-0.07	0.001	1.33	0.022	2.44**	0.041	2.28**	0.039
1.5	0.8	0.014	3.13**	0.053	0.55	0.009	-0.66	0.011	7.74***	**0.131**
1	31.22***	***0.528***	11.66***	**0.197**	7.85***	**0.133**	13.91***	**0.235**	30.66***	***0.518***
0.5	88.87***	***1.502***	28.26***	***0.478***	30.02***	***0.507***	33.51***	***0.566***	89.80***	***1.518***
0	166.74***	***2.818***	44.68***	**0.755**	57.26***	***0.968***	42.76***	**0.723**	172.58***	***2.917***
-0.5	243.35***	***4.113***	47.80***	**0.808**	89.68***	***1.516***	42.00***	**0.71**	247.06***	***4.176***
-1	303.36***	***5.128***	58.40***	**0.987**	100.39***	***1.697***	35.63***	**0.602**	319.96***	***5.408***
-1.5	339.46***	***5.738***	41.93***	**0.709**	83.98***	***1.419***	25.85***	***0.437***	363.64***	***6.147***
-2	344.30***	***5.82***	37.85***	**0.64**	53.72***	***0.908***	13.04***	**0.22**	371.33***	***6.277***
-2.5	319.92***	***5.408***	35.52***	**0.6**	29.98***	***0.507***	5.48***	0.093	343.08***	***5.799***
-3	299.18***	***5.057***	36.43***	**0.616**	17.01***	**0.288**	2.22**	0.038	308.75***	***5.219***

Note. T=true ability TS=test statistics EF=effect size
p*<.05 p**<.01 p***<.001
. Bold number indicates small effect size (.1~.29);
Bold+Italicized (in red) number indicates medium effect size (.3~.59);
Bold+underlined number indicates large effect size (.6~.89);
Bold+Italicized+underlined number indicates extremely large effect size (>.9).

Table 11

Multivariate tests with interaction (RIS_set)

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.201	8.071[a]	2.000	64.000	.001
	Wilks' Lambda	.799	8.071[a]	2.000	64.000	.001
	Hotelling's Trace	.252	8.071[a]	2.000	64.000	.001
	Roy's Largest Root	.252	8.071[a]	2.000	64.000	.001
ExpoCtrl	Pillai's Trace	.202	1.829	8.000	130.000	.077
	Wilks' Lambda	.807	1.811[a]	8.000	128.000	.081
	Hotelling's Trace	.228	1.792	8.000	126.000	.085
	Roy's Largest Root	.151	2.452[b]	4.000	65.000	.055
DIFF	Pillai's Trace	.991	3386.886[a]	2.000	64.000	.000
	Wilks' Lambda	.009	3386.886[a]	2.000	64.000	.000
	Hotelling's Trace	105.840	3386.886[a]	2.000	64.000	.000
	Roy's Largest Root	105.840	3386.886[a]	2.000	64.000	.000
ExpoCtrl * DIFF	Pillai's Trace	.080	.681	8.000	130.000	.708
	Wilks' Lambda	.920	.685[a]	8.000	128.000	.704
	Hotelling's Trace	.087	.688	8.000	126.000	.701
	Roy's Largest Root	.087	1.410[b]	4.000	65.000	.240

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+ExpoCtrl+DIFF+ExpoCtrl * DIFF

Table 12

Multivariate tests with main effect (RIS_set)

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.209	8.958[a]	2.000	68.000	.000
	Wilks' Lambda	.791	8.958[a]	2.000	68.000	.000
	Hotelling's Trace	.263	8.958[a]	2.000	68.000	.000
	Roy's Largest Root	.263	8.958[a]	2.000	68.000	.000
ExpoCtrl	Pillai's Trace	.238	2.335	8.000	138.000	.022
	Wilks' Lambda	.772	2.348[a]	8.000	136.000	.021
	Hotelling's Trace	.282	2.360	8.000	134.000	.021
	Roy's Largest Root	.220	3.796[b]	4.000	69.000	.008
DIFF	Pillai's Trace	.991	3903.254[a]	2.000	68.000	.000
	Wilks' Lambda	.009	3903.254[a]	2.000	68.000	.000
	Hotelling's Trace	114.802	3903.254[a]	2.000	68.000	.000
	Roy's Largest Root	114.802	3903.254[a]	2.000	68.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+ExpoCtrl+DIFF

Table 13

Tests of between-subjects effects (RIS_set)

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	MSE	7.916[a]	5	1.583	19.800	.000
	BIAS	11.089[b]	5	2.218	1580.598	.000
Intercept	MSE	.979	1	.979	12.238	.001
	BIAS	.004	1	.004	2.825	.097
ExpoCtrl	MSE	.378	4	.094	1.181	.327
	BIAS	.021	4	.005	3.796	.008
DIFF	MSE	7.564	1	7.564	94.602	.000
	BIAS	11.017	1	11.017	7851.747	.000
Error	MSE	5.517	69	.080		
	BIAS	.097	69	.001		
Total	MSE	20.128	75			
	BIAS	14.070	75			
Corrected Total	MSE	13.433	74			
	BIAS	11.186	74			

a. R Squared = .589 (Adjusted R Squared = .560)

b. R Squared = .991 (Adjusted R Squared = .991)

Table 14

Pairwise comparisons for BIAS (RIS_set)

	RS	SH	AS	BAS	WO
RS				.042*	
SH				.042*	
AS				.042*	
BAS					-.042*
WO					

*. The mean difference is significant at the .05 level.
Blank indicates the difference is not statistically significant.

Table 15

Multivariate tests with interaction (MI_set)

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.281	12.520[a]	2.000	64.000	.000
	Wilks' Lambda	.719	12.520[a]	2.000	64.000	.000
	Hotelling's Trace	.391	12.520[a]	2.000	64.000	.000
	Roy's Largest Root	.391	12.520[a]	2.000	64.000	.000
ExpoCtrl * DIFF	Pillai's Trace	.762	10.008	8.000	130.000	.000
	Wilks' Lambda	.282	14.137[a]	8.000	128.000	.000
	Hotelling's Trace	2.391	18.830	8.000	126.000	.000
	Roy's Largest Root	2.324	37.761[b]	4.000	65.000	.000
ExpoCtrl	Pillai's Trace	.179	1.595	8.000	130.000	.132
	Wilks' Lambda	.823	1.640[a]	8.000	128.000	.120
	Hotelling's Trace	.214	1.683	8.000	126.000	.109
	Roy's Largest Root	.205	3.330[b]	4.000	65.000	.015
DIFF	Pillai's Trace	.805	132.414[a]	2.000	64.000	.000
	Wilks' Lambda	.195	132.414[a]	2.000	64.000	.000
	Hotelling's Trace	4.138	132.414[a]	2.000	64.000	.000
	Roy's Largest Root	4.138	132.414[a]	2.000	64.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+ExpoCtrl * DIFF+ExpoCtrl+DIFF

Table 16

Tests of between-subjects effects (MI_set)

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	MSE	905.913[a]	9	100.657	37.986	.000
	BIAS	80.453[b]	9	8.939	42.899	.000
Intercept	MSE	29.155	1	29.155	11.003	.001
	BIAS	5.297	1	5.297	25.419	.000
ExpoCtrl * DIFF	MSE	376.769	4	94.192	35.547	.000
	BIAS	22.206	4	5.551	26.641	.000
ExpoCtrl	MSE	11.454	4	2.864	1.081	.373
	BIAS	2.703	4	.676	3.242	.017
DIFF	MSE	519.210	1	519.210	195.942	.000
	BIAS	51.664	1	51.664	247.935	.000
Error	MSE	172.238	65	2.650		
	BIAS	13.545	65	.208		
Total	MSE	1369.791	75			
	BIAS	131.813	75			
Corrected Total	MSE	1078.151	74			
	BIAS	93.998	74			

a. R Squared = .840 (Adjusted R Squared = .818)

b. R Squared = .856 (Adjusted R Squared = .836)

Table 17

Pairwise comparisons for BIAS (MI_set)

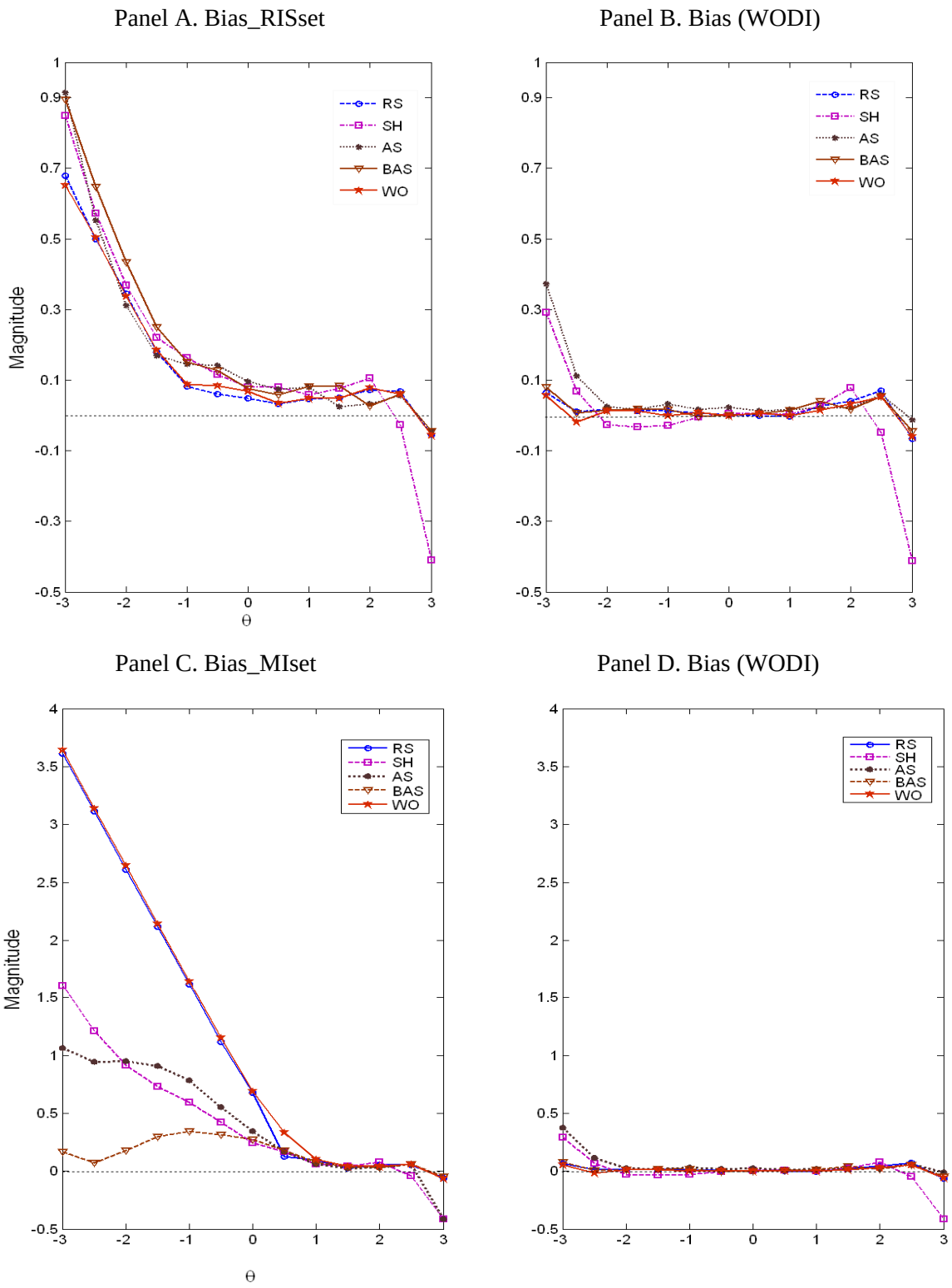
	RS	SH	AS	BAS	WO
RS		.88*	.988*	1.286*	
SH					-.916*
AS					-1.023*
BAS					-1.321*
WO					

Table 18

Pairwise comparison for MSE (MI_set)

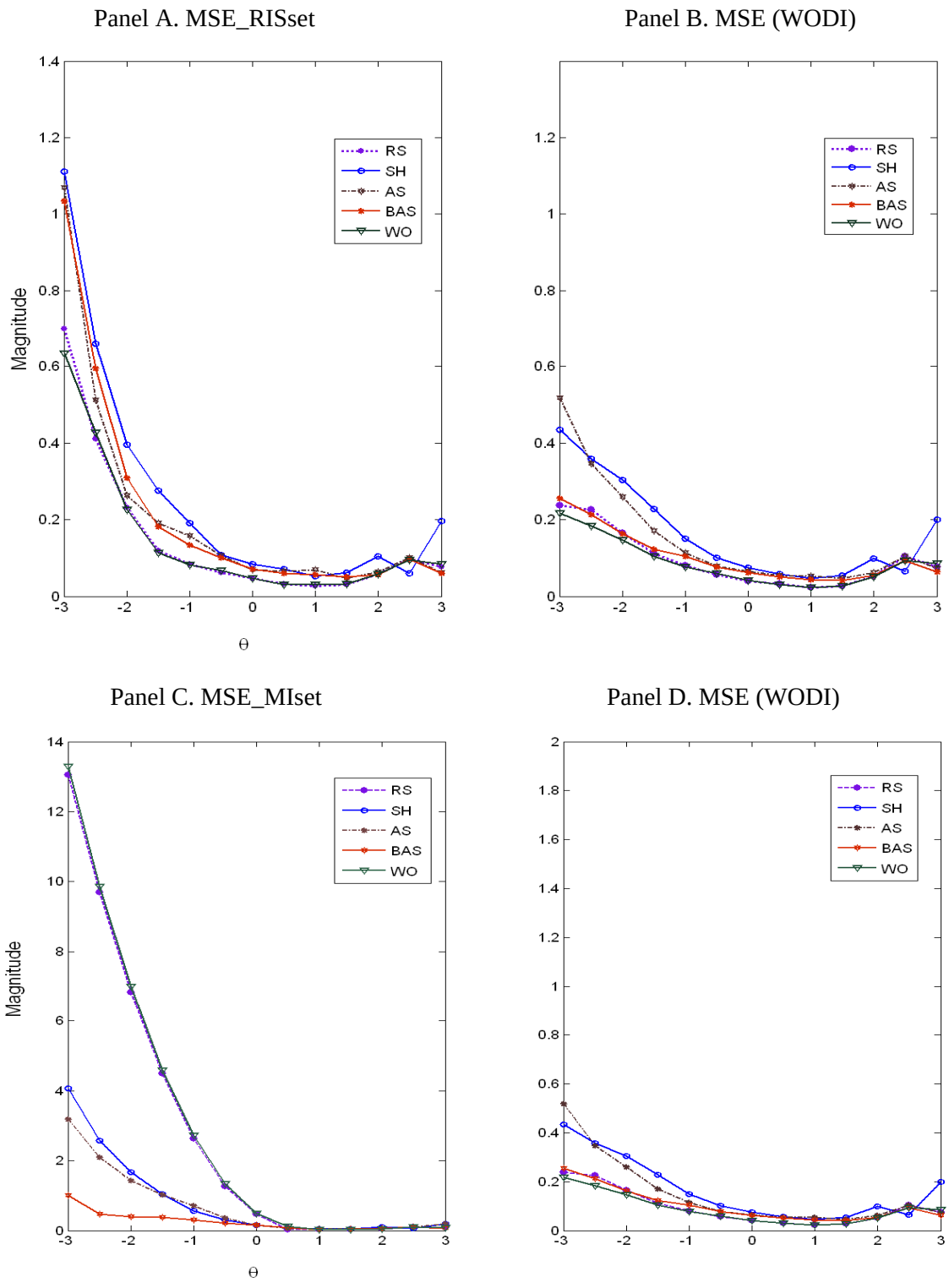
	RS	SH	AS	BAS	WO
RS		2.93*	3.456*	3.996*	
SH					-3.036*
AS					-3.562*
BAS					-4.102*
WO					

Figure 1. A comparison of biases given by different sets of disclosed items across different item exposure control procedures



Note. Panels B and D are the same. WODI=without disclosed items

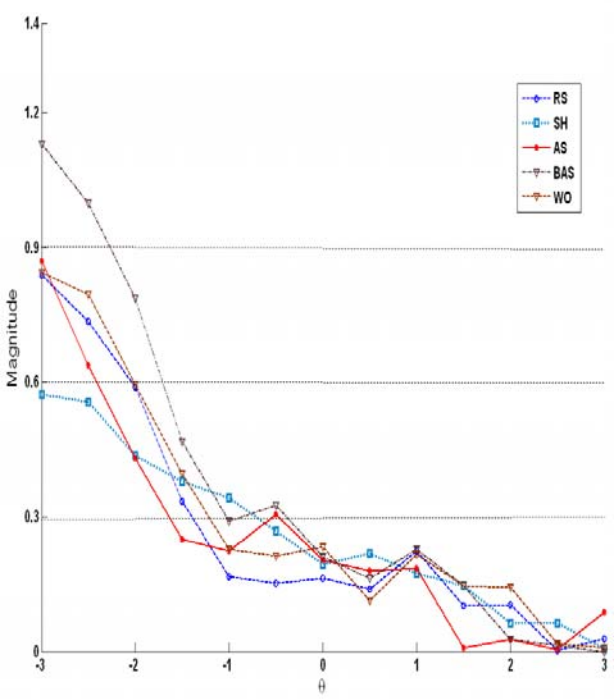
Figure 2. A comparison of MSEs given by different sets of disclosed items across different item exposure control procedures



Note. Panels B and D are the same. WODI=without disclosed items

Figure 3. Graphical representation of effect sizes from paired-sample t-tests

Panel A. RIS_set



Panel B. MI_set